Big Data in Finance

July 14, 2018

Mao Ye
University of Illinois, Urbana-Champaign and NBER

Sentiment
Completeness
Financial
Causality
Order Book
Ecommerce Transaction
Data Flows
Processors
Integration
Banking
Mortgage
News
Retail
Interpretable
Industrial
Clustering

Parallel
Variety
Web Searches
Volume
Velocity
Storage
Unstructured
MapReduce
Debit Card
Accounting Data
Three Aspects of Big Data

• Large size

• High dimension
  – A large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format
  – Satellite images, social media, and credit card transactions
Roadmap

• Large size

• High dimension
  – A large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format

• Big data motivate new economic theories
Roadmap

• Large size

• High dimension
  – Large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format

• Big data motivate new economic theories
Small vs. Large Data

- Smaller datasets often involve selection processes from larger datasets
  - Smaller sample size
  - Fewer variables
  - Aggregations of economic activity
  - Snapshot of economic activity

- Are there sample selection biases in smaller datasets?
Size of Trade and Quote Data (TAQ)

- NYSE, NASDAQ, and regional exchange listed securities
- All trades and quotes reported to the consolidated tape
## Order Level Data

<table>
<thead>
<tr>
<th>Type</th>
<th>Timestamp (nanoseconds)</th>
<th>Order Reference Number</th>
<th>Buy/Sell</th>
<th>Shares</th>
<th>Stock</th>
<th>Price</th>
<th>Original Order Reference Number</th>
<th>Market Participant ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>53435.759668667</td>
<td>335531633</td>
<td>S</td>
<td>300</td>
<td>EWA</td>
<td>19.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>40607.031257842</td>
<td>168914198</td>
<td>B</td>
<td>100</td>
<td>NOK</td>
<td>9.38</td>
<td></td>
<td>UBSS</td>
</tr>
<tr>
<td>U</td>
<td>53520.367102587</td>
<td>336529765</td>
<td></td>
<td>300</td>
<td></td>
<td>19.45</td>
<td>335531633</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>53676.740300677</td>
<td>336529765</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>57603.003717685</td>
<td>625843333</td>
<td></td>
<td>100</td>
<td></td>
<td>32.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>53676.638521222</td>
<td>336529765</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>53676.740851701</td>
<td>336529765</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **A**: Add order anonymously
- **F**: Add order with market participant ID
- **U**: Update: replace old order with a new order
- **E**: Order execution
- **C**: Order executed with price message
- **X**: Partial cancellation
- **D**: Order deletion
Research Question

• Are there selection biases in TAQ data?

• Method: Compare TAQ data with order level data
  – A large dataset and a larger dataset
Solution: High Performance Computing

- Extreme Science and Engineering Discovery Environment (XSEDE)

- First parallel: Day by day
  - Reduce data size to less than 100 gigabytes per day

- Second parallel: Among stocks
  - Each daily file contains 7,000 stocks
    - Some stocks, such as AAPL, are more actively traded than others
    - Divide stock files into equally-sized bundles
Selection Bias Led by Regulations

• Previous regulations: No need to report trades less than 100 shares (odd lots)
  – Rationale: Odd lots are from small retail traders

• Consequence: Odd lots are missing from TAQ data

• O’Hara, Yao, and Ye (2014) find:
  – 25% of trades are unreported in 2011
  – More trades are missing for high-priced stocks
    • Google: 53% of trades, 23% of volume
    • Apple: 38% of trades, 14% of volume
Are Odd Lots from Retail Traders?

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Symbol</th>
<th>Hour</th>
<th>Minute</th>
<th>Second</th>
<th>Millisecond</th>
<th>Shares</th>
<th>Buy/Sell</th>
<th>Price</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>20</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>2</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>10</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>109</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>110</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>30</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>111</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>107</td>
<td>3</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>112</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>47</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>113</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>80</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>114</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>80</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>210</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>5</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>211</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>25</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>212</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>50</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
<tr>
<td>213</td>
<td>AAPL</td>
<td>13</td>
<td>59</td>
<td>1</td>
<td>110</td>
<td>12</td>
<td>S</td>
<td>125.00</td>
<td>HN</td>
</tr>
</tbody>
</table>
Machines Challenge Regulations

• Computers can reduce large orders to small odd lots
  – Benefit: Hide information
  – Odd lots are more informed than trades greater than or equal to 100 shares

• Policy impact: Regulators reduce report threshold from 100 shares to 1 share
100 Share Cutoff + $5,000 Cutoff

- Lee and Radhakrishna (2000) method
  - Use trades less than $5,000 in TAQ data as a proxy for individual trades

- We find $5,000 cutoff leads to data truncation based on stock price
  - Zero individual trading for any stock with a price higher than $50
  - Truncation does not depend directly on the market share of odd lots
  - O’Hara, Yao, and Ye (2014) estimate the magnitude of truncation based on CRSP
Individual Trading (False Zeroes)

- Up to 15% percent of stocks are truncated
- Those stocks represent up to 70% market cap
- Generate mechanical patterns of retail trading
  - Retail traders trade less in dot-com bubble period
Size Challenges

Techniques
• XSEDE helps to overcome size challenges

Economic insights
• Open question for policy
  – Many regulations were designed for humans
  – Should regulations be revised for machines?

• Are there selection biases in other “small” datasets?
  – Can larger datasets lead to different results?
Roadmap

- Large size

- High dimension
  - Large number of variables relative to the sample size

- Complex structure
  - Not in traditional row-column format

- Big data motivate new economic theories
Does Machine Learning Capture Any Economic Signal?

• Firms that use machine-learning techniques to make investment decisions, such as Renaissance Technologies and Two Sigma Investments, operate at timescales ranging “anywhere from a few minutes to a few months.”

• Chinco, Clark-Joseph, and Ye (2017)
  – Examine this question at minute-by-minute horizon
High Dimensional Challenges

- Basic idea: Use lagged stock returns to forecast $r_{n,t+1}$

- Data: One-minute returns of other ($\approx 2,000$) NYSE-listed stocks

- OLS requires at least 2,000 observations (six trading days)
  - Too many RHS variables for OLS
  - Hard-to-capture signals that are unexpected and short-lived

- We use machine learning techniques to overcome this dimensional challenge
Differences in Approaches

Traditional Approaches

Step 1: Use economic reasoning to select $x$.

Step 2: Use statistical approach to estimate $x$’s quality
  • Sorting
  • Linear regression

Machine Learning Techniques
  • Use statistical approach to select and estimate $x$
  • Handle a large number of $x$ variables
  • More flexible functional form
Machine Learning Techniques

• Two common features
  – Cross-validation: Methods are evaluated using out-of-sample predictions
    • Less emphasis on causal inference
    • Belloni, Chernozhukov, and Hansen (2014); Athey and Imbens (2017); Mullainathan and Spies (2017)
  – Regularization: Penalty for complex models to avoid overfitting

• Two variations
  – Functional form: Linear, regression trees, or neural networks
  – The type of regularization

• Example: Chinco, Clark-Joseph, and Ye (2017) use LASSO
Functional Form

OLS

\[
\min_\beta \left\{ \frac{1}{2 \cdot T} \cdot \sum_{t=1}^{T} \left( r_{n,t} - \beta_0 - \sum_{n'=1}^{N} \beta_{n'} \cdot r_{n',t-1} \right)^2 \right\}
\]
LASSO is OLS with a penalty function:

\[
\min_\beta \left\{ \frac{1}{2 \cdot T} \sum_{t=1}^{T} \left( r_{n,t} - \beta_0 - \sum_{n'=1}^{N} \beta_{n'} \cdot r_{n',t-1} \right)^2 + \lambda \sum_{n'=1}^{N} |\beta_{n'}| \right\}
\]

- Variable selection
  - Variables are normalized before regression
  - Small OLS coefficients are set to 0
Cross Validation

We search the best penalty parameter $\lambda$ through k-fold cross validation

- We choose $k = 10$

1. For every $k = 1, \ldots, 10$, use the $k^{th}$ portion as the testing sample, the rest as training sample

2. Use training sample to calculate the LASSO estimator

3. Use testing sample to calculate the mean squared error:

   $$Q(k, \lambda) = \sum_{t=1}^{T} \left( r_{n,t} - \beta_0(k, \lambda) - \sum_{n' = 1}^{N} \tilde{\beta}_{n'}(k, \lambda) \cdot r_{n', t-1} \right)^2$$

4. Repeat steps 2–3 ten times to get the average:

   $$\bar{Q}(\lambda) = \frac{1}{10} \sum_{k=1}^{10} Q(k, \lambda)$$

5. Pick the $\lambda$ with the best overall performance:

   $$\hat{\lambda}_{min} = \arg \min_{\lambda} \bar{Q}(\lambda)$$

Forecast-Implied Performance Net of Trading Costs

Annualized Sharpe Ratios

<table>
<thead>
<tr>
<th></th>
<th>S&amp;P 500</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.123</td>
<td>1.791</td>
</tr>
</tbody>
</table>

LASSO-Implied Strategy Abnormal Returns [%/yr]

<table>
<thead>
<tr>
<th>LASSO-Implied Strategy</th>
<th>α</th>
<th>Mkt</th>
<th>HmL</th>
<th>SmB</th>
<th>Mom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td>2.709</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Factor Model</td>
<td>2.713</td>
<td>0.004</td>
<td>-0.004</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.002)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td></td>
</tr>
<tr>
<td>4-Factor Model</td>
<td>2.707</td>
<td>0.005</td>
<td>-0.004</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.002)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
</tr>
</tbody>
</table>
Choice of Predictors

• Unexpected
  – LASSO ignores well-known weekly or monthly predictors
  – Reason: LASSO typically ignores predictors weaker than 2.5% per month

• Short-lived: 95% of LASSO predictors last less than 14.2 minutes

• Sparse: LASSO uses only 12.7 predictors on average

• LASSO is more likely to pick a stock as a predictor before its news announcements
  – Even if we use the millisecond news feeds like RavenPack
Machine Learning vs. News

• Big data incorporate information faster than news announcements

• Writing news articles takes time, especially for unscheduled events
  – The difference between public information and news

• Empirical evidence
  – LASSO is more likely to pick a stock as a predictor before unscheduled news
  – LASSO is more likely to pick a stock as a predictor in the same minute as scheduled news
Three Open Questions

• Other horizons
  – Feng, Giglio, and Xiu (2018); Freyberger, Neuhier, and Weber (2018); Han, He, Rapach, and Zhou (2018)

• Other regularizations
  \[
  \min_{\beta} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left( r_{n,t} - \beta_0 - \sum_{n'=1}^{N} \beta_{n'} \cdot r_{n',t-1} \right)^2 + \lambda \sum_{n'=1}^{N} \beta_{n'}^2 \right\}
  \]

• Other functional forms
  – Capture important nonlinearities and interactions
    • Gu, Kelly, and Xiu (2018)
  – Titanic example (Varian, 2014)
    • Logistic regression predicts “age barely matters for survival rate”
    • Regression tree predicts survival rate is very high for passengers less than 8.5 years old
High Dimensional Challenges

• Techniques
  – Machine learning techniques deal with high dimensional data

• Economic insights
  – Determining economic interpretations is a higher hurdle
Roadmap

• Large size

• High dimension
  – A large number of variables relative to the sample size

• Complex structure
  – Not in traditional row-column format

• Big data motivate new economic theories
Types of Unstructured Data (Kolanovic and Rajesh, 2017)

- Generated by individuals
  - Social media posts, news, product reviews, web search records, mobile apps, personal pictures/videos/audios

- Generated by business transactions and government filings
  - Supermarket scanner data, SEC filings

- Generated by sensors
  - Satellite images, weather and pollution sensors
Example: Twitter Data
Two Challenges

• Techniques: How to extract information from unstructured data?
  – One solution: Find a data vendor
    • Many vendors transfer unstructured data to structured data (e.g., RavenPack)
    • A comprehensive list of 500 alternative data vendors
    • J.P. Morgan’s Big Data and AI Strategies (2017)
  – Another solution: Interdisciplinary collaboration

• Economics: Do unstructured data generate unique measures of economic activity?
  – More challenging

• Example: Da, Nitesh, Xu, and Ye (2017)
Two Challenges

- Techniques: How to extract information from unstructured data?
  - One solution: Find a data vendor
    - Many vendors transfer unstructured data to structured data. E.g. Ravenpack
    - A comprehensive list of 500 alternative data vendors
    - J.P. Morgan’s Big Data and AI Strategies (2017)
  - Another solution: interdisciplinary collaboration

- Economics: Do unstructured data generate unique measures of economic activity?
  - More challenging

- Example: Da, Nitesh, Xu, and Ye (2017)
Unique Measures from Big Data

- **Information diffusion**
  - Word-of-mouth communication: No direct measure without big data

- **Two traditional solutions**
  - Proxies: Physical proximity (Hong, Kubik, and Stein, 2005; Ivkovich and Weisbenner, 2007; Brown et al., 2008) and common schooling (Cohen, Frazzini, and Malloy, 2008)
  - Criminal investigations (Rantala, 2015; Ahern, 2016)

- **Big data solution**
  - Measure information diffusion using tweets and retweets
Information Diffusion through Retweets

Zhi has 10,000 followers

@Zhi: Twitter data are unstructured ...

Nitesh has 100,000 followers

@Nitesh @Zhi: Twitter ...

Jian has 5,000 followers

@Jian @Nitesh @Zhi: Twitter …
Useful Fields in the Original Tweet

twitter_public_stream.20140128-220104.json:

```
"created_at": "Wed Jan 29 21:14:11 +0000 2014",
"id": 428637220338425856,
"id_str": "428637220338425856",
"text": "Facebook earnings: Q4 EPS $0.31 ex-items v. $0.27 estimate; revenues $2.59 billion v. $2.33 billion estimate - @CNBC http://t.co/sNqdBtfzyv",
"source": "\"\"",
"truncated": false,
"in_reply_to_status_id": null,
"in_reply_to_status_id_str": null,
"in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
"in_reply_to_screen_name": null,
"user": {
  "id": 6017542,
  "id_str": "6017542",
  "name": "Breaking News",
  "screen_name": "BreakingNews",
  "location": "Global",
  "url": "http://www.breakingnews.com/about/mobile",
  "description": "Introducing our new iOS app and http://www.breakingnews.com that lets you control the breaking news you want to see.",
  "protected": false,
  "followers_count": 6483805,
  "friends_count": 475,
  "listed_count": 85853,
  "created_at": "Sun May 13 23:06:45 +0000 2007",
  "favourites_count": 81,
  "utc_offset": -18000,
  "time_zone": "Eastern Time (US & Canada)",
  "geo_enabled": false,
  "verified": true,
  "statuses_count": 82721,
  "lang": "en",
  "contributors_enabled": false,
  "is_translation_enabled": true,
  "profile_background_color": "E6E6FA",
  "profile_background_image_url": "http://a0.twimg.com/profile_background_images/661943965/2eu2ntwqt6ereyuum38.png",
  "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/661943965/2eu2ntwqt6ereyuum38.png",
  "profile_background_tile": "false",
  "profile_image_url": "http://pbs.twimg.com/profile_images/378800000706301008/53d967d7656bd591e7e1fcdff47e0b_normal.png",
  "profile_image_url_https": "https://pbs.twimg.com/profile_images/378800000706301008/53d967d7656bd591e7e1fcdff47e0b_normal.png",
  "profile_banner_url": "https://pbs.twimg.com/profile_banners/6017542/1383589267",
  "profile_link_color": "CC0000",
  "profile_sidebar_border_color": "FFFFFF",
  "profile_sidebar_fill_color": "F3F3F3",
  "profile_text_color": "333333",
  "profile_use_background_image": true,
  "default_profile": false,
  "default_profile_image": false,
  "following": null,
  "follow_request_sent": null,
  "notifications": null,
  "geo": null,
  "coordinates": null,
  "place": null,
  "contributors": null,
  "retweet_count": 0,
  "favorite_count": 0,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "urls": []
  }
}
```
One Retweet

facebook_earnings_change_20140128-220104.json: 
"created_at": "Wed Jan 29 21:14:33 +0000 2014", "id": 428637311690366976, "text": "Facebook earnings: Q4 EPS $0.31 ex-items v. $0.27 estimate; revenues $2.59 billion v. $2.33 billion estimate - @CNBC http://t.co/sNgDQfhyzy", "source": "u003ca href="http://www.breakingnews.com" rel="nofollow" u003c/a u003e, "truncated": false, "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_screen_name": null, "user": {"id": 6017542, "id_str": "6017542", "name": "Breaking News", "screen_name": "BreakingNews", "location": "Global", "url": "http://www.breakingnews.com/about/mobile", "description": "Introducing our new iOS app and http://www.breakingnews.com that lets you control the breaking news you want to..."
Speed of Information Diffusion

Number of Accounts Reached

- p5
- p50
- p95
Social media can spread stale news
  - When someone retweets news, it is already stale
    - Stale: Ten minutes after the initial release from a news outlet
  - Retail traders still respond
    - Create temporal price pressures
    - Prices first overshoot then revert to the next day

Smart traders should trade against stale news
  - Profit opportunity: Sell after stale good news and quickly buy back
Machines vs. Humans?

• Reversion speed in our sample period (2013–2014) is much faster than reported in Tetlock (2011)

• Open question: Are smart traders machines?

• Broader questions
  – Do machines trade against human behavioral biases?
  – Are markets more efficient due to the rise of machines?
Structure Challenges

- **Techniques**
  - Find an alternative data vendor
  - Work with experts in other fields

- **Economic insights**
  - Unstructured data create unique measures of economic activity
  - Unstructured data help financial economists to test economic theory
Roadmap

- Large size

- High dimension
  - A large number of variables relative to the sample size

- Complex structure
  - Not in traditional row-column format

- Big data motivate new economic theories
What Drives the Arms Race in Speed?

Standard Walrasian equilibrium
- Continuous price

Reality: Price is discrete
- Tick size
- Minimum price increment imposed by SEC Rule 612
Liquidity Supplier: Limit Order Submitter

Limit order
An offer to buy or sell

- $100.02
- $100.01
- $100.00

Toni:
An offer to buy 100 shares at $100.00

$99.99
Liquidity Demander: Market Order Submitter

Limit order
An offer to buy or sell

Market order
Accepts the limit order

Toni:
An offer to buy 100 shares at $100.00

Mao:
Sells 100 shares at $100.00
Execution Priority for Liquidity Suppliers

1. Price: First priority
   - Better-priced limit orders execute first
     • Limit sells at lower prices
     • Limit buys at higher prices

2. Time: Second priority
   - At the same price: first come, first served
Identification: ETF Splits/Reverse Splits

• Hypothesis: High-frequency traders (HFTs) provide more liquidity for low-priced securities
  – One cent tick size is more binding

• Treatment group: ETFs that split/reverse split
  – Splits decrease price
  – Reverse splits increase price

• Control group: ETFs track the same index but do not split/reverse split

• Tens of TBs of trading data (Yao and Ye, 2018)
  – Supercomputer helps to analyze 64 splits/reverse splits in four years
Price vs. Speed Competition

$100.04  🧡🧡🧡 dhcp
$100.03  🧡🧡🧡 🧡 dhcp
$100.02  🧡🧡🧡 🧡 🧡 🧡 $50.02
$100.01  🧡🧡🧡 🧡 🧡 $50.01
$100.00  🧡🧡🧡 $50.00

HFT: 🧡
Non-HFT: 🧡
Puzzles

• Who are these non-HFTs?

• Why do they quote better prices than HFTs?

• Analysis of big data → new theory → new analysis of big data
Two Types of Traders

- HFTs: Computers
- Humans
The Third Type

- HFTs: Computers
- Half human, half computer
- Humans
Half Human, Half Computer
Buy-side Algorithmic Traders (BATs)

• BATs: Half human and half computer

• Humans (e.g., portfolio managers) make investment decisions

• Algorithms execute orders for portfolio managers
  – Demand or provide liquidity to minimize transaction costs

• BATs are faster than humans

• BATs are slower than HFTs
  – No need to be as fast as micro or nanoseconds

- Continuous time, continuous price, and **two** types of traders

- HFTs: Continually monitor the market
  - Supply or demand liquidity if the expected profit is above 0

- Non-HFTs arrive with an inelastic need to buy/sell one share
  - Arrival intensity $\lambda_I$
  - Only demand liquidity

- Security value $v_t$ evolves as a compound Poisson process
  - $v_t$ is public information
  - Intensity of the jump event: $\lambda_J$
  - Size of the jump: $d$ or $-d$
Revenue to Supply Liquidity to a Non-HFT

- **Sell at ask**
  - $V_t + \frac{s}{2}$
  - $V_t$
  - $V_t - \frac{s}{2}$

- **HFTs**
  - HFT’s profit is $\frac{s}{2}$

- **Non-HFTs**
Cost of Being Sniped

\[ V_t + \frac{s}{2} \]

- Liquidity provider loses money
- Snipers make money

\[ V_t + d \]

- HFTs turn into snipers
- Non-HFTs

Sell at ask

\[ V_t - \frac{s}{2} \]
Adding BATs (Li, Wang, and Ye, 2018)

\[ V_t + \frac{s^*}{2} \]
\[ V_t - \frac{s^*}{2} \]

Sell at ask

Buy at bid

Humans

Non-HFTs

BATs

HFTs
Continuous Price: BATs always Supply Liquidity

\[ V_t + \frac{s^*}{2} \]

\[ V_t - \frac{s^*}{2} \]

Transaction cost +\( \varepsilon \)

Sell at ask

Buy at bid

To buy

HFTs

Humans

BATs

Transaction cost +\( \varepsilon \)

Sell at ask

Buy at bid

\[ V_t + \varepsilon \]

\[ V_t \]
Machine-Machine Interactions

- BATs always provide liquidity
  - Lower opportunity costs for providing liquidity
  - BATs have to buy or sell

- HFTs’ strategy to sell
  - Offer to sell at $V_t + \frac{s^*}{2}$
    - Limit sell orders are more likely to be executed when $v_t$ jumps upward
  - Accept BATs’ offer at $V_t + \varepsilon$
    - Immediate execution with no sniping risk

- Machine-machine interactions blur the definition of liquidity provision
  - BATs arrive first
  - HFTs immediately respond
Predictions and Policy Implications

• Li, Wang, and Ye (2018) explain several existing puzzles
  – Non-HFTs quote better prices due to lower opportunity costs

• New predictions after adding discrete size: Four types of equilibria
  – Theory works well to predict who provides liquidity and when
  – Machine-machine interactions provide a clean environment to test theories

• Policy implication: SEC’s tick size pilot program
  – Increase the tick size to 5 cents for 1,200 pilot stocks
  – Implication from the model: A large tick size would fuel speed competition
Financial Market Ecosystem

Underexplored Territories

Microseconds
High-frequency Traders

3 Months
13F Data

• Paucity of studies on traders who are slower than HFTs but faster than a quarter
  • BATs operate at timescales of milliseconds or seconds
  • Traders who use machine-learning techniques operate at timescales of “anywhere from a few minutes to a few months.”
Our Solution: Wavelet Estimator

- Aggregate each stock’s trading-volume data to the one-minute timescale from TAQ data

- Decompose each stock’s trading-volume variance into timescale-specific components with the wavelet-variance estimator
  – Chinco and Ye (2018)
Intuition

Simulated Minute Trading-Volume Time Series

What happens in different horizons?

Low frequency

Median frequency

High Frequency
Wavelet Decomposition
Big Data Research Strategy
Conclusion: Big Data Challenges and Opportunities

Techniques
- High-performance computing mitigates the size challenges
- Machine learning alleviates the high dimensional challenges
- Alternative data vendors or interdisciplinary collaborations mitigate the structure challenges

Big data open doors for new research questions
- Document new empirical regularities and inform public policy
- Motivate us to find economic interpretations of new data
- Create unique measures to test theories and motivate us to construct new theories